Fourier-Net: Fast Image Registration with Band-Limited Deformation

Xi Jia[1], **Joseph Bartlett**[1,2], **Wei Chen**[1], **Siyang Song**[3], **Tianyang Zhang**[1], **Xinxing Cheng**[1], **Wenqi Lu**[4], **Zhaowen Qiu**[5†], **Jinming Duan**[1,6†]

[1] School of Computer Science, University of Birmingham, UK
[2] Department of Biomedical Engineering, University of Melbourne, Australia
[3] Department of Computer Science and Technology, University of Cambridge, UK
[4] Department of Computer Science, University of Warwick, UK
[5] Institute of Information Computer Engineering, Northeast Forestry University, China
[6] Alan Turing Institute, UK

Abstract

Unsupervised image registration commonly adopts U-Net style networks to predict dense displacement fields in the full-resolution spatial domain. For high-resolution volumetric image data, this process is however resource-intensive and time-consuming. To tackle this problem, we propose the Fourier-Net, replacing the expansive path in a U-Net style network with a parameter-free model-driven decoder. Specifically, instead of our Fourier-Net learning to output a full-resolution displacement field in the spatial domain, we learn its low-dimensional representation in a band-limited Fourier domain. This representation is then decoded by our devised model-driven decoder (consisting of a zero padding layer and an inverse discrete Fourier transform layer) to the dense, full-resolution displacement field in the spatial domain. These changes allow our unsupervised Fourier-Net to contain fewer parameters and computational operations, resulting in faster inference speeds. Fourier-Net is then evaluated on two public 3D brain datasets against various state-of-the-art approaches. For example, when compared to a recent transformer-based method, named TransMorph, our Fourier-Net, which only uses 2.2% of its parameters and 6.66% of the multiply-add operations, achieves a 0.5% higher Dice score and an 11.48 times faster inference speed. Code is available at https://github.com/xi-jia/Fourier-Net.

1 Introduction

Medical image registration aims to learn a spatial deformation that identifies the correspondence between a moving image and a fixed image, which is a fundamental step in many medical image analysis applications such as longitudinal studies, population modeling, and statistical atlases (Sotiras, Davatzikos, and Paragios 2013).

Iterative optimization techniques such as FFD (Rueckert et al. 1999), Demons (Vercauteren et al. 2009), ANTS (Avants et al. 2011), Flash (Zhang and Fletcher 2019) and ADMM (Thorley et al. 2021) have been applied to deformable image registration. However, such optimization-based approaches require elaborate hyper-parameter tuning for each image pair, and iteration towards an optimal deformation is very time-consuming, thus limiting their applications in real-time and large-scale volumetric registration.

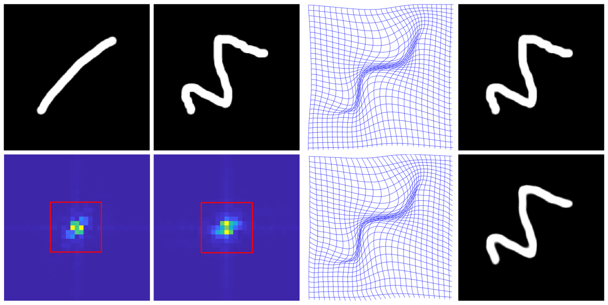

Figure 1: From left to right in the 1st row: moving image, fixed image, deformation grid, and warped moving image. From left to right in the 2nd row: DFT of horizontal displacement field, DFT of vertical displacement field, reconstructed deformation grid from Fourier coefficients only in band-limited region (marked by red rectangles), and the warped moving image by reconstructed deformation.

In recent years, deep learning based approaches have burgeoned in the field of medical image registration (Hering et al. 2022). Their success has been largely driven by their exceptionally fast inference speeds. The most effective methods, such as VoxelMorph (Balakrishnan et al. 2019), usually adopt a U-Net style architecture to estimate dense, spatial deformation fields. They only require one forward propagation during inference, and thus can register images several orders of magnitudes faster than traditional iterative methods. Following the success of VoxelMorph, a large number of deep learning based approaches have been proposed for various registration tasks (Zhang 2018; Zhao et al. 2019; Mok and Chung 2020b; Jia et al. 2021; Kim et al. 2021; Chen et al. 2021a; Jia et al. 2022). These models either use multiple U-Net style networks in a cascaded way or replace basic convolution blocks in VoxelMorph with more sophisticated ones such as swin-transformers (Chen et al. 2021a) to boost registration performance. However, these changes rapidly increase the number of network parameters and multiply-add operations (mult-adds), sacrificing training and inference efficiency altogether.

For U-Net-based registration models, we argue 1) that it may be unnecessary to include the whole expansive path of U-Net backbone and 2) that the training and inference effi-

† Corresponding author.

ciency of such networks can be further improved by learning a low-dimensional representation of displacement field in a band-limited Fourier domain. Our arguments are based on our observation in Figure 1, where we notice it is sufficient to reconstruct an accurate full-resolution deformation (the third image of the second row in Figure 1) by using only a small number of coefficients in the band-limited Fourier domain. Inspired by this insight, we propose an end-to-end unsupervised approach that learns only a low-dimensional representation of displacement field in the band-limited Fourier domain. We term our approach the Fourier-Net.

By removing several layers in the expansive path of a U-Net style architecture, our Fourier-Net outputs only a small patch that stores low-frequency coefficients of displacement field in the Fourier domain. We then propose to directly apply a model-driven decoder to recover the full-resolution spatial displacement field from these low-frequency Fourier coefficients. This model-driven decoder contains a zero-padding layer that broadcasts complex-valued low-frequency signals into a full-resolution complex-valued map. The inverse discrete Fourier transform (iDFT) is then applied to this map to obtain the full-resolution spatial displacement field. Both zero-padding and iDFT layers are parameter-free and therefore fast. On top of Fourier-Net, we propose a diffeomorphic variant, termed Fourier-Net-Diff. This network first estimates a stationary velocity field, followed by squaring and scaling layers (Dalca et al. 2018) to encourage the output deformation to be diffeomorphic.

2 Related Works

Unsupervised approaches can be based on either iterative optimization or learning. Iterative methods are prohibitively slow, especially when the images to be registered are of a high-dimensional form, such as 3D volumes. Over the past decades, many works have been proposed to accelerate such methods. (Ashburner 2007) used a stationary velocity field (SVF) representation (Legouhy et al. 2019), and proposed a fast algorithm DARTEL for image registration which computes the resulting deformation by using scaling and squaring from the SVF. Another fast approach for image registration is Demons (Vercauteren et al. 2009), which imposes smoothness on displacement fields by incorporating inexpensive Gaussian convolutions into its iterative process. Hernandez (Hernandez 2018) reformulated the Stokes-LDDMM variational problem used in (Mang and Biros 2015) in the domain of band-limited non-stationary vector fields and utilized GPUs to parallelize their methods. (Zhang and Fletcher 2019) developed the Fourier-approximated Lie algebras for shooting (Flash) for fast diffeomorphic image registration, where they proposed to speed up the solution of the Euler-Poincaré differential (EPDiff) equation used to compute deformations from velocity fields in the band-limited Fourier domain.

On the other hand, deep learning methods based on convolutional neural networks have been employed to overcome slow registration speeds. Among them, U-Net style networks have been proven to be an effective tool to learn deformations between pairwise images (Balakrishnan et al. 2019; Zhang 2018; Mok and Chung 2020b; Kim et al. 2021). While their registration performance is comparable with iterative methods, their inference can be orders of magnitude faster. RC-Net (Zhao et al. 2019) and VR-Net (Jia et al. 2021) cascaded multiple U-Net style networks to improve the registration performance, but their speed is relatively slow. Very recently, approaches, such as ViT-V- Net (Chen et al. 2021b) and TransMorph (Chen et al. 2021a), which combine vision transformers and U-Nets have achieved promising registration performance, but they involve much more computational operations and are therefore slow. Another group of network-based image registration techniques (De Vos et al. 2019; Qiu et al. 2021) is to estimate a grid of B-Spline control points with regular spacing, which is then interpolated based on cubic B-Spline basis functions (Rueckert et al. 1999; Duan et al. 2019). By estimating fewer control points, these networks perform fast predictions, but currently are less accurate.

Supervised approaches are also studied in medical image registration. However, they have several pitfalls: 1) it is generally hard to provide human-annotated ground truth deformations for supervision; and 2) if trained using numerical solutions of other iterative methods, the performance of these supervised registration methods may be limited by iterative methods. Yang et al. proposed Quicksilver (Yang et al. 2017) which is a supervised encoder-decoder network and trained using the initial momentum of LDDMM as the supervision signal. Wang et al. extended Flash (Zhang and Fletcher 2019) to DeepFlash (Wang and Zhang 2020) in a learning framework in lieu of iterative optimization. Compared to Flash, DeepFlash accelerates the computation of initial velocity fields but needs to solve a PDE (i.e., EPDiff equation) in the Fourier domain so as to recover the full-resolution deformation in the spatial domain, which can be slow. The fact that DeepFlash requires the numerical solutions of Flash (Zhang and Fletcher 2019) as training data attributes to lower registration performance than Flash.

Although DeepFlash also learns a low-dimensional band-limited representation, it differs from our Fourier-Net in four aspects, which we reckon our novel contributions to this area. First, DeepFlash is a supervised method that requires ground truth velocity fields calculated from Flash prior to training, whilst Fourier-Net is a simple and effective unsupervised method thanks to our proposed model-driven decoder. Second, DeepFlash is a multi-step method whose network's output requires an additional PDE algorithm (Zhang and Fletcher 2019) to compute final full-resolution spatial deformations, whilst Fourier-Net is a holistic model that can be trained and used in an end-to-end manner. Third, DeepFlash needs two individual convolutional networks to estimate real and imaginary signals in the band-limited Fourier domain, whilst Fourier-Net uses only one single network directly mapping image pairs to a reduced-resolution displacement field without the need of complex-valued operations. Lastly, DeepFlash is essentially an extension of Flash and it is difficult for the method to benefit from vast amounts of data, whilst Fourier-Net is flexible and can easily learn from large-scale datasets. Due to these, our Fourier-Net outperforms DeepFlash (as well as Flash) by a significant margin in terms of both accuracy and speed.

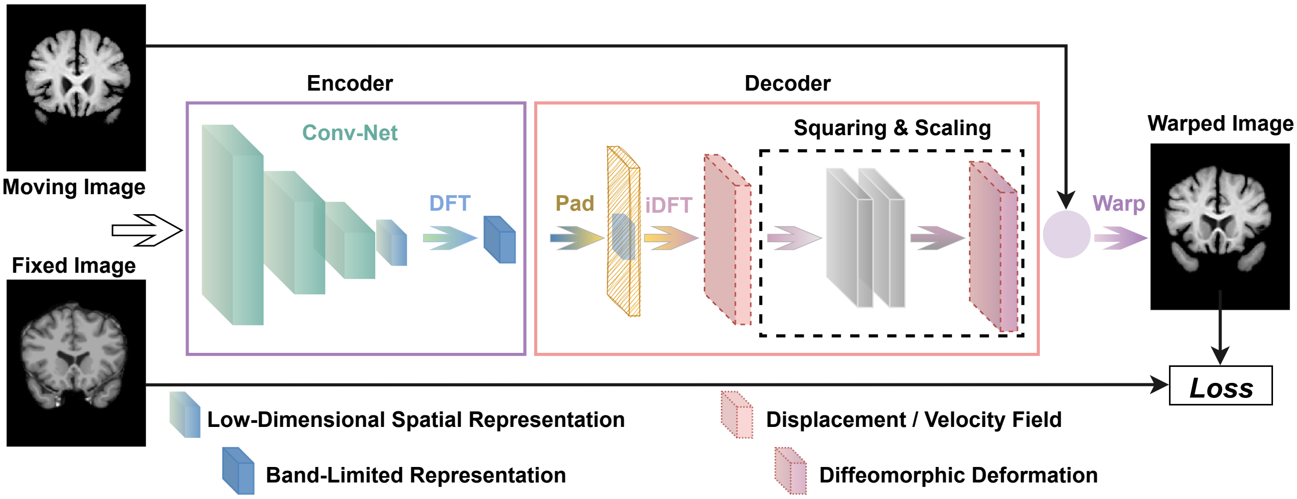
Moving Image
Fixed Image
Encoder
Conv-Net
DFT
Decoder
Squaring & Scaling
Pad
iDFT
Warp
Warped Image
Loss
Low-Dimensional Spatial Representation
Band-Limited Representation
Displacement / Velocity Field
Diffeomorphic Deformation

Figure 2: Architecture of our end-to-end Fourier-Net. It contains 1) a convolutional encoder that first produces a low-dimensional representation of displacement or velocity field, followed by an embedded discrete Fourier transformation (DFT) layer to map this low-dimensional representation into the band-limited Fourier domain; 2) a parameter-free model-driven decoder that adopts a zero-padding layer, an inverse DFT (iDFT) layer, and seven optional squaring and scaling layers to reconstruct the displacement field or deformation into the full-resolution spatial domain from its band-limited Fourier domain; 3) a warping layer to deform the moving image; and 4) a loss function that includes a similarity term and a regularization term.

3 Methodology

As illustrated in Figure 2, in Fourier-Net, the encoder first takes a pair of spatial images as input, and encodes them to a low-dimensional representation of displacement field (or velocity field if diffeomorphisms are imposed) in the band-limited Fourier domain. Then the decoder brings the displacement field (or velocity field) from the band-limited Fourier domain to the spatial domain, and ensures that they have the same spatial size as the input image pair. Next, the optional squaring and scaling layers are used to encourage a diffeomorphism in final deformations. Finally, by minimizing the loss function, an accurate deformation can be estimated, with which the warping layer deforms the moving image to be similar to the fixed image.

3.1 Encoder

The encoder aims to learn a displacement or velocity field in the band-limited Fourier domain. Intuitively, this may require convolutions to be able to handle complex-valued numbers. One may directly use complex convolutional networks (Trabelsi et al. 2017), they are suitable when both input and output are complex values, but complex-valued operations sacrifice computational efficiency. Instead, DeepFlash (Wang and Zhang 2020) tackles this problem by first converting input image pairs to the Fourier domain and then using two individual real-valued convolutional networks to learn the real and imaginary signals separately. Such an approach increase the training and inference cost (as listed in Table 1). Since our Fourier-Net estimates displacement fields in the band-limited Fourier domain from spatial images (inputs are real values but outputs are complex values), these approaches may not be well suited to our application.

To bridge the domain gap between real-valued spatial images and complex-valued band-limited displacement fields without increasing complexity, we propose to embed a DFT layer at the end of the convolutional network in the encoder. This is a simple and effective way to produce complex-valued band-limited displacement fields without the network being able to handle complex values itself. Let us denote the moving image as I_0, the fixed image as I_1, the convolutional network as CNN with the parameters $\mathbf{\Theta}$, the DFT as $\mathcal{F}$, the full-resolution spatial displacement field as ϕ, and the complex band-limited displacement field as $\mathbb{B}_\phi$. In this case, our encoder can be defined as $\mathbb{B}_\phi = \mathcal{F}(\text{CNN}(I_0, I_1; \mathbf{\Theta}))$, resulting in a compact, efficient implementation as compared to DeepFlash and other complex convolutional networks. On the other hand, we also notice from our experiments (Table 1) that it is difficult to regress $\mathbb{B}_\phi$ directly from I_0 and I_1 within a single CNN, i.e., $\mathbb{B}_\phi = \text{CNN}(I_0, I_1; \mathbf{\Theta})$. We believe the reason being that: if the CNN directly learns a band-limited displacement field, it needs to go through two domains altogether: first mapping the spatial images to the spatial displacement field and then mapping this displacement field into its band-limited Fourier domain. In this case, the domain gap is too big. Our network however only needs to go through one domain and then DFT handles the second domain. By doing so, Fourier-Net is efficient and easy to learn. An illustration of this idea is given in Figure 3.

So far, we have given an intuitive explanation of how the encoder in our network learns. Here we discuss their mathematical relationship between the low-dimensional spatial displacement field $\mathbb{S}_\phi = \text{CNN}(I_0, I_1; \mathbf{\Theta})$, its band-limited representation $\mathbb{B}_\phi$, as well as the displacement field ϕ (coming after the decoder) in the full-resolution spatial domain. For simplicity, we use a 2D displacement field as an example and the formulations below can be easily extended to 3D

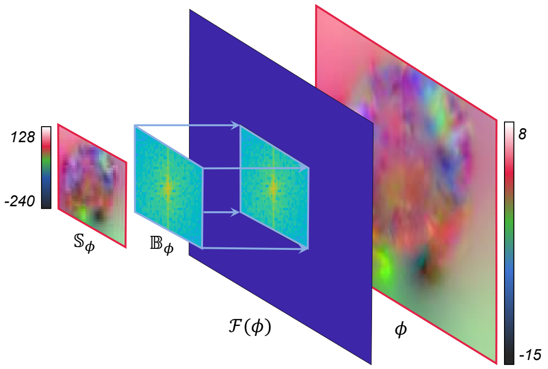

Figure 3: Connection between low-dimensional spatial displacement field $\mathbb{S}_\phi$, band-limited Fourier coefficients $\mathbb{B}_\phi$, full-resolution Fourier coefficients $\mathcal{F}(\phi)$ by zero-padding $\mathbb{B}_\phi$, and full-resolution displacement field ϕ by taking iDFT of $\mathcal{F}(\phi)$.

cases. A general discrete Fourier transform used on ϕ can be defined as follows:

$$[\mathcal{F}(\phi)]_{k,l} = \sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\phi_{i,j}e^{-\sqrt{-1}\left(\frac{2\pi k}{M}i+\frac{2\pi l}{N}j\right)}, \quad (1)$$

where ϕ is of size $M \times N$, $i \in [0, M-1]$ and $j \in [0, N-1]$ are the discrete indices in the spatial domain, and $k \in [0, M-1]$ and $l \in [0, N-1]$ are the discrete indices in the frequency domain.

In our Fourier-Net, ϕ is actually a low-pass filtered displacement field. If we define a $M \times N$ sized sampling mask $\mathcal{D}$ whose entries are zeros if they are on the positions of high-frequency signals in ϕ and ones if they are on the low-frequency positions. With $\mathcal{D}$, we can recover the displacement field ϕ from Eq. (1)

$$\phi_{i,j} = \frac{1}{MN}\sum_{k=0}^{M-1}\sum_{l=0}^{N-1}\mathcal{D}_{k,l}[\mathcal{F}(\phi)]_{k,l}e^{\sqrt{-1}\left(\frac{2\pi i}{M}k+\frac{2\pi j}{N}l\right)}. \quad (2)$$

If we shift all low-frequency signals of the displacement field to a center patch of size $\frac{M}{a} \times \frac{N}{b}$ ($\frac{M}{a}, \frac{N}{b}, a = 2Z_a, b = 2Z_b, Z_a, Z_b \in \mathbb{Z}^+$), center-crop the patch (denoted by $\mathbb{B}_\phi$), and then perform the iDFT on this patch, we obtain $\mathbb{S}_\phi$ in Eq. (3)

$$[\mathbb{S}_\phi]_{\hat{i},\hat{j}} = \frac{ab}{MN}\sum_{\hat{k}=0}^{\frac{M}{a}-1}\sum_{\hat{l}=0}^{\frac{N}{b}-1}[\mathbb{B}_\phi]_{\hat{k},\hat{l}}e^{\sqrt{-1}\left(\frac{2\pi a\hat{i}}{M}\hat{k}+\frac{2\pi b\hat{j}}{N}\hat{l}\right)}, \quad (3)$$

where $\hat{i} \in [0, \frac{M}{a}-1]$ and $\hat{j} \in [0, \frac{N}{b}-1]$ are the indices in the spatial domain, and $\hat{k} \in [0, \frac{M}{a}-1]$ and $\hat{l} \in [0, \frac{N}{b}-1]$ are the indices in the frequency domain. Note that $\mathbb{S}_\phi$ is a low-dimensional spatial representation of ϕ and we are interested in their mathematical connection. Another note is that $\mathbb{S}_\phi$ actually contains all the information of its band-limited Fourier coefficients in $\mathbb{B}_\phi$. As such, we do not need the network to learn the coefficients in $\mathbb{B}_\phi$ and instead only to learn its real-valued coefficients in $\mathbb{S}_\phi$.

Since most of entries ($\frac{a\times b-1}{a\times b}$%) in $\mathcal{F}(\phi)$ are zeros, and the values of rest entries are exactly the same as in $\mathbb{B}_\phi$, we can conclude that $\mathbb{S}_\phi$ contains all the information ϕ can provide, and their mathematical connection is

$$[\mathbb{S}_\phi]_{\hat{i},\hat{j}} = ab \times \phi_{a\hat{i},b\hat{j}}. \quad (4)$$

With this derivation, we show that we can actually recover a low-dimensional spatial representation $\mathbb{S}_\phi$ from its full-resolution spatial displacement field ϕ, as long as they have the same low-frequency coefficients $\mathbb{B}_\phi$. This essentially proves that there exists a unique mapping function between $\mathbb{S}_\phi$ and ϕ and that it is reasonable to use a network to learn $\mathbb{S}_\phi$ directly from image pairs.

3.2 Model-Driven Decoder

The proposed decoder consists of a zero-padding layer, an iDFT layer, and an optional squaring and scaling module.

The output from the encoder is a band-limited representation $\mathbb{B}_\phi$. To recover the full-resolution displacement field ϕ in the spatial domain, we first pad the patch $\mathbb{B}_\phi$ containing mostly low-frequency signals to the original image resolution with zero values (i.e., $\mathcal{F}(\phi)$). We then feed the zero-padded complex-valued coefficients $\mathcal{F}(\phi)$ to an iDFT layer consisting of two steps: shifting the Fourier coefficients from centers to corners and then applying the standard iDFT to convert them into the spatial domain. The output from Fourier-Net is thus a full-resolution spatial displacement field. Both padding and iDFT layers are differentiable and therefore Fourier-Net can be optimized via standard back-propagation. We note that our proposed decoder is a parameter-free module that is driven by knowledge instead of learning and therefore fast.

We also propose a diffeomorphic variant of Fourier-Net which we term Fourier-Net-Diff. A diffeomorphic deformation is defined as a smooth and invertible deformation, and in Fourier-Net-Diff we need an extra squaring and squaring module for the purpose. The output of the iDFT layer can be regarded as a stationary velocity field denoted by $\boldsymbol{v}$ instead of the displacement field ϕ. In group theory, $\boldsymbol{v}$ is a member of Lie algebra, and we can exponentiate this stationary velocity field (i.e., $Exp(\boldsymbol{v})$) to obtain a diffeomorphic deformation. In this paper, we use seven scaling and squaring layers (Ashburner 2007; Dalca et al. 2018) to impose such a diffeomorphism.

3.3 Warping Layer and Loss Functions

After the model-driven decoder, we obtain a full-resolution displacement field (or a diffeomorphic deformation) for the input image pair. We then deform the moving image using a warping layer to produce the warped moving image, which is then used to calculate the loss. We implement 2D and 3D spatial warping layers based on linear interpolation as in (Jaderberg et al. 2015) and (Balakrishnan et al. 2019).

We adopt an unsupervised loss which is computed from the moving image I_1, the fixed image I_0, and the predicted displacement field $\boldsymbol{\phi}$ or velocity field $\boldsymbol{v}$. The training objective of our Fourier-Net is $\mathcal{L}(\boldsymbol{\Theta}) = \min_{\boldsymbol{\Theta}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{Sim}(I_1^i \circ (\boldsymbol{\phi}_i(\boldsymbol{\Theta}) + \mathrm{Id}) - I_0^i) + \frac{\lambda}{N} \sum_{i=1}^{N} \|\nabla \boldsymbol{\phi}_i(\boldsymbol{\Theta})\|_2^2$, where N is the number of training pairs, $\boldsymbol{\Theta}$ is the network parameters to be learned, Id is the identity grid, $\circ$ is the warping operator, and ∇ is the first order gradient implemented using finite differences (Lu et al. 2016; Duan et al. 2016). The first term $\mathcal{L}_{Sim}$ defines the similarity between warped moving images and fixed images, and the second term defines the smoothness of displacement fields. Here λ is a hyper-parameter balancing the two terms. As for Fourier-Net-Diff, the training loss is defined as $\mathcal{L}(\boldsymbol{\Theta}) = \min_{\boldsymbol{\Theta}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{Sim}(I_1^i \circ Exp(\boldsymbol{v}_i(\boldsymbol{\Theta})) - I_0^i) + \frac{\lambda}{N} \sum_{i=1}^{N} \|\nabla \boldsymbol{v}_i(\boldsymbol{\Theta})\|_2^2$. $\mathcal{L}_{Sim}$ can be either mean squared error (MSE) or normalized cross-correlation (NCC), which we clarify in our experiments.

4 Experiments

4.1 Datasets

OASIS-1 dataset (Marcus et al. 2007) consists of a cross-sectional collection of T1-weighted brain MRI scans from 416 subjects. In experiments, we use the pre-processed OASIS data[1] provided by (Hoopes et al. 2021) to perform subject-to-subject brain registration, in which all 414 MRI scans are bias-corrected, skull-stripped, aligned, and cropped to the size of 160 × 192 × 224. Automated segmentation masks from FreeSurfer are provided for evaluation of registration accuracy. This dataset also has 414 2D slices and marks extracted from their corresponding 3D volumes. We randomly split this 2D dataset into 201, 12, and 201 images for training, validation, and test. After pairing, we end up with 40200 (201×200), 22 ([12-1]×2), and 400 ([201-1]×2) image pairs for training, validation, and test, respectively.

IXI dataset[2] contains nearly 600 MRI scans from healthy subjects. In experiments, we use the pre-processed IXI data provided by (Chen et al. 2021a) to perform atlas-based brain registration. The atlas is generated by (Kim et al. 2021). There are in total 576 160×192×224 3D brain MRI volumes in this dataset. The dataset is split into 403 for training, 58 for validation, and 115 for testing. There is no pairing step for this dataset as it is an atlas-to-subject registration task.

4.2 Implementation Details

We implement our Fourier-Net using PyTorch, where training is optimized using Adam with a fixed learning rate of 0.0001. We tune built-in hyper-parameters on a held-out validation set. Specifically, we use MSE to train both 2D and 3D OASIS for 10 and 1000 epochs, respectively, and λ in $\mathcal{L}(\boldsymbol{\Theta})$ is set to 0.01. For 3D OASIS, an additional Dice loss is used with its weight being set to 1. On 3D IXI, we train the models with NCC loss for 1000 epochs with $\lambda = 5$. All deep models are trained with an Nvidia A100 GPU.

The CNN in Fourier-Net has 6 convolutional blocks. The initial 4 blocks contain 2 convolutional layers in each block. The first layer maintains the same spatial resolution as inputs, while the second layer performs a down-sampling with a stride of 2 and then doubles the number of feature channels. In the last 2 blocks, each contains a fractional convolutional layer and 2 convolutional layers. The fractional layer performs an up-sampling with a stride of 2, and the convolutional layers halve the number of feature channels. The kernel size in all convolutional layers is 3×3×3. Each convolution is followed by a PReLU activation except the last sub-layer, which does not have any activation layer and contains 2 or 3 kernels for 2D or 3D registration, respectively. The initial number of kernels in the first convolutional layer is set to C. For example, the spatial resolution of input images changes from 160×192×224×2 to 80×96×112×C → 40×48×56×2C → 20×24×28×4C → 10×12×14×8C → 20×24×28×4C → 40×48×56×3 after each block. We experiment C=8, 16, and 48, which define small Fourier-Net$_{\mathbf{S}}$, Fourier-Net, and large Fourier-Net$_{\mathbf{L}}$, respectively. Though the output of our Fourier-Net is set to 40×48×56, the resolution is not constrained and one can customize the CNN architecture to produce a band-limited representation with any resolution. To adapt Fourier-Net onto 2D images, we directly change all 3D kernels to 2D.

Patch	DFT	SS	Dice↑	$\lvert J\rvert_{<0}\%$	MA(M)
20×24	✗	✗	.664±.040	.158±.206	891
20×24	✓	✗	.732±.042	.434±.355	679
20×24	✓	✓	.735±.037	0.0±0.0	679
40×48	✗	✗	.675±.038	.279±.257	1310
40×48	✓	✗	.756±.039	.753±.407	888
40×48	✓	✓	.756±.037	<0.0001	888
U-Net	–	✓	.762±.039	<0.0001	2190

Table 1: Ablation and parameter studies. SS denotes squaring and scaling, $|J|_{<0}\%$ is the percentage of negative values of Jacobian determinant of deformation. MA(M) refers to the number of mult-adds operations in millions.

4.3 Ablation Studies and Parameter Tuning

The first question we ask is what the most suitable resolution (i.e., patch size) of a band-limited displacement field is? A very small patch will rapidly decrease model parameters as well as training and inference time but may lead to lower performance. A very large patch could retain registration accuracy but may increase training and inference time, thus eliminating the advantages of our method.

In Table 1, we use 2D OASIS images for ablation studies and investigate the impact of different patch sizes, i.e, 20×24 and 40×48, which are respectively $\frac{1}{64}$ and $\frac{1}{16}$ of the original image size. It can be seen that the 40×48 patch improves Dice by 2% over the 20×24 patch, with only a slight difference in mult-adds operations. The Dice score of our Fourier-Net (40×48) is already close to the full-resolution U-Net backbone (last row in this Table), which however has 2.5 times mult-adds cost than our Fourier-Net (40×48).

We also prove the necessity of embedding a DFT layer in the encoder. Without this layer, our encoder is purely a CNN that has to learn complex-valued Fourier coefficients from image pairs. Following DeepFlash (Wang and Zhang 2020), we use two networks to separately compute the real and imaginary parts of these complex coefficients. As re-

[1] https://learn2reg.grand-challenge.org/Learn2Reg2021/

[2] https://brain-development.org/ixi-dataset/

Methods	Patch	Dice↑	$\|J\|_{<0}\%$	CPU
Initial	-	.544±.089	-	-
Flash	16×16	.702±.051	.033±.126	13.7
Flash	20×24	.727±.046	.205±.279	22.6
Flash	40×48	.734±.045	.049±.080	85.8
DeepFlash	16×16	.615±.055	0.0±0.0	.487
DeepFlash	20×24	.597±.066	0.0±0.0	.617
B-Spline-Diff	20×24	.737±.038	.015±.069	.012
B-Spline-Diff	40×48	.735±.040	.009±.055	.012
F-Net$_\mathbf{S}$	40×48	.748±.039	.671±.390	**.007**
F-Net-Diff$_\mathbf{S}$	40×48	.750±.038	<0.0001	.010
F-Net	40×48	.756±.039	.753±.407	.011
F-Net-Diff	40×48	.756±.037	<0.0001	.015
F-Net$_\mathbf{L}$	40×48	.759±.040	.781±.406	.037
F-Net-Diff$_\mathbf{L}$	40×48	**.761±.037**	0.0±0.0	.040

Table 2: Comparing different methods on 2D OASIS. F-Net is the abbreviation for Fourier-Net. All reported CPU runtimes (in seconds) are tested on the same machine.

Methods	Dice↑	HD95↓
Initial	.572±.053	3.831
nnU-Net (Hering et al. 2022)	.846±.016	1.500
LapIRN	.861±.015	1.514
TransMorph	.858±.014	1.494
TransMorph-Large	**.862±.014**	1.431
Fourier-Net-Diff	.843±.013	1.495
Fourier-Net	.847±.013	1.455
Fourier-Net$_\mathbf{L}$	.860±.013	**1.375**

Table 3: Performance comparison on 3D OASIS which is the MICCAI Learn2reg 2021 Task 3 validation dataset. All results are taken from the leaderboard. HD95 is the 95% Hausdorff distance, a lower value suggests a better performance.

ported in Table 1, using the DFT layer, Dice is improved by 6.8% and 8.1% for the patch sizes of 20×24 and 40×48, respectively, which validates the efficacy of such a layer. This experiment shows the superiority of our proposed band-limited representation over DeepFlash's.

We further study the impact of adding a squaring and scaling module into Fourier-Net. As shown in Table 1, this module encourages diffeomorphisms for the estimated deformation, due to the fact that it produces less percentage of negative values of Jacobian determinant of deformation.

4.4 Comparison on Inter-subject Registration

2D OASIS: In Table 2, we compare the performance of Fourier-Net with Flash (Zhang and Fletcher 2019), DeepFlash (Wang and Zhang 2020), and B-Spline-Diff (Qiu et al. 2021). We manage to compile and run Flash[3] in CPU, but its official GPU version keeps throwing *segmentation fault* errors. We report the performance of Flash on three band-limited patch sizes, and its built-in hyper-parameters are grid-searched over 252 different combinations on the whole validation set for each size. We also manage to run DeepFlash[4] with supervision from Flash's results. We train DeepFlash on all 40200 training pairs for 1000 epochs with more than 40 different combinations of hyper-parameters and report the best results. B-Spline-Diff is also trained with all training pairs using its official implementation[5].

In Table 2, all Fourier-Net variants outperform competing methods in terms of Dice. Specifically, our Fourier-Net$_\mathbf{S}$ achieves a 0.748 Dice score with 0.007 seconds inference speed per image pair. Compared to Flash using a 40×48 patch, Fourier-Net$_\mathbf{S}$ improves Dice by 1.5% and is 12,257 times faster. Though DeepFlash is much faster than Flash, we find that DeepFlash is very difficult to converge and as such achieves the lowest Dice score (0.597). Moreover, DeepFlash is not an end-to-end method, because its output (band-limited velocity field) requires an additional PDE algorithm to compute the final deformation. As such, it is much slower than deep learning methods such as ours or B-Spline-Diff (0.012 seconds per image pair on CPU). Note that the computational time is averaged on the whole test set, including the cost of loading models and images.

Note that the speed advantage of Fourier-Net on CPU decreases when we use larger models such as Fourier-Net$_\mathbf{L}$, but its performance can be boosted by 1.1% compared to Fourier-Net$_\mathbf{S}$ in terms of Dice.

We also list the percentage of negative values of Jacobian determinant of deformation for all compared methods in Table 2. Though both Flash and B-Spline-Diff are diffeomorphic approaches, neither of them produces perfect diffeomorphic deformations on this dataset. The proposed three Fourier-Net-Diff variants, however, barely generate negative Jacobian determinants and are therefore diffeomorphic.

3D OASIS: In Table 3, we further compare Fourier-Net with other methods on the MICCAI 2021 Learn2reg challenge dataset. Though Fourier-Net is slightly lower than LapRIN(Mok and Chung 2020a) in Dice, it achieves a better Hausdorff distance than LapRIN with a 0.059 improvement. If we use a larger Fourier-Net$_\mathbf{L}$, it can achieve the lowest HD95, suggesting that Fourier-Net is able to obtain comparable results on par with state-of-the-art on this dataset.

4.5 Comparison on Atlas-Based Registration

3D IXI: In Table 4, we first compare our Fourier-Net with iterative methods such as Flash and deedsBCV (Heinrich, Maier, and Handels 2015) and deep learning methods such as TransMorph-B-Spline (Chen et al. 2021a), which is a combination of TransMorph and B-Spline-Diff. Note that for Flash, 200 combinations of hyper-parameters are grid-searched using 5 randomly selected validation samples. We do not include all images in the validation set because tuning Flash on CPU can take up to *30 minutes* for each pair.

The proposed Fourier-Net achieves the highest Dice score (0.763) with 4.42s inference speed per image pair. By using less number of kernels in each layer, Fourier-Net$_\mathbf{S}$ achieves the fastest inference speed (1.92s) on CPU, which is faster than all other deep learning methods, while retaining a competitive accuracy. Furthermore, Fourier-Net$_\mathbf{S}$ outperforms

[3] https://bitbucket.org/FlashC/flashc/src/master/

[4] https://github.com/jw4hv/deepflash

[5] https://github.com/qiuhuaqi/midir

Methods	Dice↑	$\lvert J\rvert_{<0}\%$	Parameters	Mult-Adds (G)	CPU (s)	GPU (s)
Affine*	.386±.195	-	-	-	-	-
SyN* (Avants et al. 2011)	.645±.152	<0.0001	-	-	-	-
NiftyReg* (Modat et al. 2010)	.645±.167	<0.0001	-	-	-	-
LDDMM* (Beg et al. 2005)	.680±.135	<0.0001	-	-	-	-
Flash (Zhang and Fletcher 2019)	.692±.140	0.0±0.0	-	-	-	-
deedsBCV*	.733±.126	0.147±0.050	-	-	-	-
VoxelMorph-1* (Balakrishnan et al. 2019)	.728±.129	1.590±0.339	274,387	304.05	9.373	0.391
VoxelMorph-2* (Balakrishnan et al. 2019)	.732±.123	1.522±0.336	301,411	398.81	10.530	0.441
VoxelMorph-Diff*	.580±.165	<0.0001	307,878	89.67	3.691	0.418
B-Spline-Diff* (Qiu et al. 2021)	.742±.128	<0.0001	266,387	47.05	7.076	0.437
TransMorph* (Chen et al. 2021a)	.754±.124	1.579±0.328	46,771,251	657.64	22.035	0.443
TransMorph-Diff* (Chen et al. 2021a)	.594±.163	<0.0001	46,557,414	252.61	10.389	0.438
TransMorph-B-Spline* (Chen et al. 2021a)	.761±.122	<0.0001	46,806,307	425.95	18.138	0.442
Fourier-Net$_S$	.759±.132	0.009±0.008	1,050,800	43.82	**1.919**	**0.318**
Fourier-Net-Diff$_S$	.756±.130	0.0±0.0	1,050,800	43.82	6.202	0.332
Fourier-Net	**.763±.129**	0.024±0.019	4,198,352	169.07	4.423	0.342
Fourier-Net-Diff	.761±.131	0.0±0.0	4,198,352	169.07	8.679	0.345

Table 4: Performance comparison between different methods on IXI. Results of the methods labeled with $*$ are taken from TransMorph (Chen et al. 2021a), as we used the exact same data splitting and testing protocol as TransMorph. The reported runtimes of all deep methods are computed by us on the same machine and are averaged on the whole testing set.

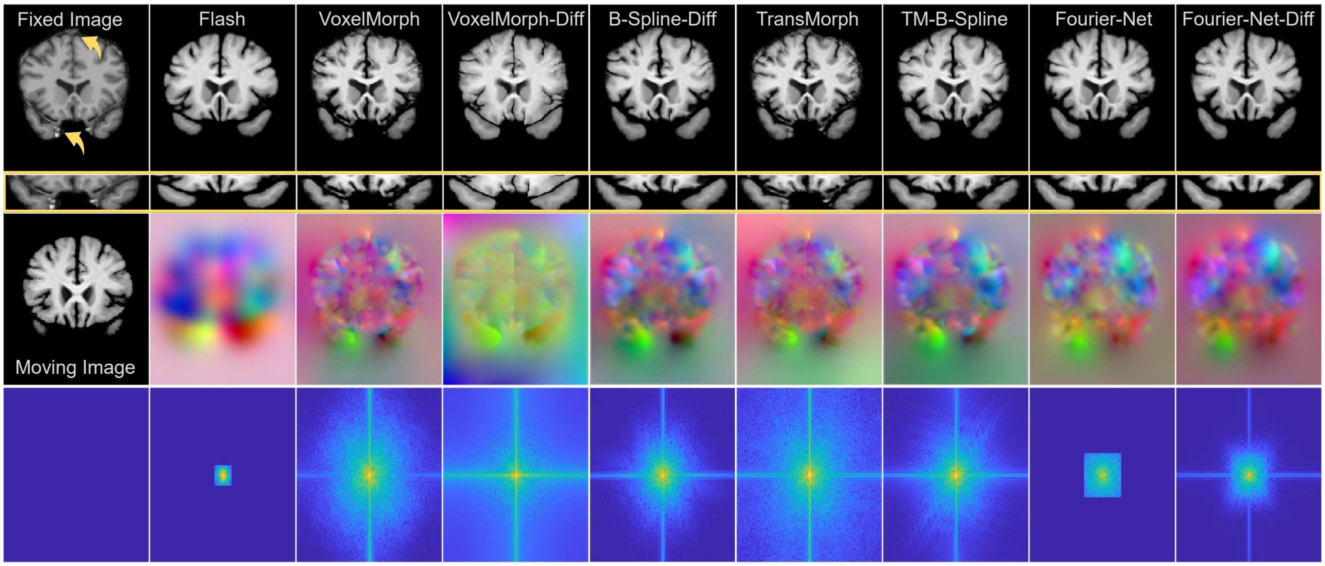

Figure 4: Visual comparison between different methods on 3D IXI. The 1st column displays a fixed image, a moving image, and a placeholder. From top to bottom rows excluding the 1st column: warped moving images (with a zoomed area in the yellow box), displacement fields, and displacement fields after DFT. Our Fourier-Net and Fourier-Net-Diff produce smoother deformations and better warped moving images (see noisy regions marked by yellow arrows and box).

TransMorph by 0.5% in Dice with only 2.2% of its parameters and 6.66% of its mult-adds. In terms of inference speed, Fourier-Net$_S$ is 11.48 times faster than TransMorph (22.035 seconds) on CPU. Finally, Table 4 (3rd column) indicates that Fourier-Net-Diff barely generates any folding and thus effectively preserves diffeomorphisms.

From Figure 4, we can observe that Flash's deformation also has no foldings, but it over-smoothes its displacement field, resulting in a less accurate warping. Figure 4 (last row) shows that only Flash and Fourier-Net produce strictly band-limited Fourier coefficients, and that the deformation of Fourier-Net-Diff is no longer band-limited due to the use of the squaring and scaling layers.

5 Conclusion

In this paper, we propose to learn a low-dimensional representation of displacement/velocity field in the band-limited Fourier domain. Experimental results on two brain datasets show that our Fourier-Net is more efficient than state-of-the-art methods in terms of speed while retaining a comparative performance in terms of registration accuracy.

Acknowledgements

The computations described in this research were performed using the Baskerville Tier 2 HPC service. Baskerville was funded by the EPSRC and UKRI (EP/T022221/1 and EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham. Xi Jia is partially supported by the Chinese Scholarship Council.

References

Ashburner, J. 2007. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1): 95–113.

Avants, B. B.; Tustison, N. J.; Song, G.; Cook, P. A.; Klein, A.; and Gee, J. C. 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*, 54(3): 2033–2044.

Balakrishnan, G.; Zhao, A.; Sabuncu, M. R.; Guttag, J.; and Dalca, A. V. 2019. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Transactions on Medical Imaging*, 38(8): 1788–1800.

Beg, M. F.; Miller, M. I.; Trouvé, A.; and Younes, L. 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2): 139–157.

Chen, J.; Frey, E. C.; He, Y.; Segars, W. P.; Li, Y.; and Du, Y. 2021a. TransMorph: Transformer for unsupervised medical image registration. *arXiv preprint arXiv:2111.10480*.

Chen, J.; He, Y.; Frey, E. C.; Li, Y.; and Du, Y. 2021b. ViT-V-Net: Vision Transformer for Unsupervised Volumetric Medical Image Registration. *arXiv preprint arXiv:2104.06468*.

Dalca, A. V.; Balakrishnan, G.; Guttag, J.; and Sabuncu, M. R. 2018. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 729–738. Springer.

De Vos, B. D.; Berendsen, F. F.; Viergever, M. A.; Sokooti, H.; Staring, M.; and Išgum, I. 2019. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52: 128–143.

Duan, J.; Bello, G.; Schlemper, J.; Bai, W.; Dawes, T. J.; Biffi, C.; de Marvao, A.; Doumoud, G.; O'Regan, D. P.; and Rueckert, D. 2019. Automatic 3D bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach. *IEEE transactions on medical imaging*, 38(9): 2151–2164.

Duan, J.; Qiu, Z.; Lu, W.; Wang, G.; Pan, Z.; and Bai, L. 2016. An edge-weighted second order variational model for image decomposition. *Digital Signal Processing*, 49: 162–181.

Heinrich, M. P.; Maier, O.; and Handels, H. 2015. Multi-modal Multi-Atlas Segmentation using Discrete Optimisation and Self-Similarities. *VISCERAL Challenge@ ISBI*, 1390: 27.

Hering, A.; Hansen, L.; Mok, T. C.; Chung, A. C.; Siebert, H.; Häger, S.; Lange, A.; Kuckertz, S.; Heldmann, S.; Shao, W.; et al. 2022. Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*.

Hernandez, M. 2018. Band-limited stokes large deformation diffeomorphic metric mapping. *IEEE Journal of Biomedical and Health Informatics*, 23(1): 362–373.

Hoopes, A.; Hoffmann, M.; Fischl, B.; Guttag, J.; and Dalca, A. V. 2021. Hypermorph: Amortized hyperparameter learning for image registration. In *International Conference on Information Processing in Medical Imaging*, 3–17. Springer.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, 2017–2025.

Jia, X.; Bartlett, J.; Zhang, T.; Lu, W.; Qiu, Z.; and Duan, J. 2022. U-Net vs Transformer: Is U-Net Outdated in Medical Image Registration? *arXiv preprint arXiv:2208.04939*.

Jia, X.; Thorley, A.; Chen, W.; Qiu, H.; Shen, L.; Styles, I. B.; Chang, H. J.; Leonardis, A.; De Marvao, A.; O'Regan, D. P.; et al. 2021. Learning a Model-Driven Variational Network for Deformable Image Registration. *IEEE Transactions on Medical Imaging*, 41(1): 199–212.

Kim, B.; Kim, D. H.; Park, S. H.; Kim, J.; Lee, J.-G.; and Ye, J. C. 2021. CycleMorph: cycle consistent unsupervised deformable image registration. *Medical Image Analysis*, 71: 102036.

Legouhy, A.; Commowick, O.; Rousseau, F.; and Barillot, C. 2019. Unbiased longitudinal brain atlas creation using robust linear registration and log-Euclidean framework for diffeomorphisms. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 1038–1041. IEEE.

Lu, W.; Duan, J.; Qiu, Z.; Pan, Z.; Liu, R. W.; and Bai, L. 2016. Implementation of high-order variational models made easy for image processing. *Mathematical Methods in the Applied Sciences*, 39(14): 4208–4233.

Mang, A.; and Biros, G. 2015. An inexact Newton–Krylov algorithm for constrained diffeomorphic image registration. *SIAM journal on imaging sciences*, 8(2): 1030–1069.

Marcus, D. S.; Wang, T. H.; Parker, J.; Csernansky, J. G.; Morris, J. C.; and Buckner, R. L. 2007. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9): 1498–1507.

Modat, M.; Ridgway, G. R.; Taylor, Z. A.; Lehmann, M.; Barnes, J.; Hawkes, D. J.; Fox, N. C.; and Ourselin, S. 2010. Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine*, 98(3): 278–284.

Mok, T. C.; and Chung, A. 2020a. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 211–221. Springer.

Mok, T. C.; and Chung, A. C. 2020b. Fast Symmetric Diffeomorphic Image Registration with Convolutional Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qiu, H.; Qin, C.; Schuh, A.; Hammernik, K.; and Rueckert, D. 2021. Learning Diffeomorphic and Modality-invariant Registration using B-splines. In *Medical Imaging with Deep Learning*.

Rueckert, D.; Sonoda, L. I.; Hayes, C.; Hill, D. L.; Leach, M. O.; and Hawkes, D. J. 1999. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8): 712–721.

Sotiras, A.; Davatzikos, C.; and Paragios, N. 2013. Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, 32(7): 1153–1190.

Thorley, A.; Jia, X.; Chang, H. J.; Liu, B.; Bunting, K.; Stoll, V.; de Marvao, A.; O'Regan, D. P.; Gkoutos, G.; Kotecha, D.; et al. 2021. Nesterov Accelerated ADMM for Fast Diffeomorphic Image Registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 150–160. Springer.

Trabelsi, C.; Bilaniuk, O.; Zhang, Y.; Serdyuk, D.; Subramanian, S.; Santos, J. F.; Mehri, S.; Rostamzadeh, N.; Bengio, Y.; and Pal, C. J. 2017. Deep complex networks. *arXiv preprint arXiv:1705.09792*.

Vercauteren, T.; Pennec, X.; Perchant, A.; and Ayache, N. 2009. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1): S61–S72.

Wang, J.; and Zhang, M. 2020. DeepFlash: An efficient network for learning-based medical image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4444–4452.

Yang, X.; Kwitt, R.; Styner, M.; and Niethammer, M. 2017. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage*, 158: 378–396.

Zhang, J. 2018. Inverse-consistent deep networks for unsupervised deformable image registration. *arXiv preprint arXiv:1809.03443*.

Zhang, M.; and Fletcher, P. T. 2019. Fast diffeomorphic image registration via fourier-approximated lie algebras. *International Journal of Computer Vision*, 127(1): 61–73.

Zhao, S.; Dong, Y.; Chang, E. I.-C.; and Xu, Y. 2019. Recursive Cascaded Networks for Unsupervised Medical Image Registration. In *The IEEE International Conference on Computer Vision (ICCV)*.